Published in: Literary and Linguistic Computing Vol. 28 (2013) No. 4 pp. 563-575

Automatic recognition of speech, thought, and writing representation in German narrative texts

Annelen Brunner
Institut für deutsche Sprache, Mannheim, Germany

Abstract

This article presents the main results of a project, which explored ways to recognize and classify a narrative feature—speech, thought, and writing representation (ST&WR)—automatically, using surface information and methods of computational linguistics. The task was to detect and distinguish four types—direct, free indirect, indirect, and reported ST&WR—in a corpus of manually annotated German narrative texts. Rule-based as well as machine-learning methods were tested and compared. The results were best for recognizing direct ST&WR (best F1 score: 0.87), followed by indirect (0.71), reported (0.58), and finally free indirect ST&WR (0.40). The rule-based approach worked best for ST&WR types with clear patterns, like indirect and marked direct ST&WR, and often gave the most accurate results. Machine learning was most successful for types without clear indicators, like free indirect ST&WR, and proved more stable. When looking at the percentage of ST&WR in a text, the results of machine-learning methods always correlated best with the results of manual annotation. Creating a union or intersection of the results of the two approaches did not lead to striking improvements. A stricter definition of ST&WR, which excluded borderline cases, made the task harder and led to worse results for both approaches.

Correspondence:
Annelen Brunner, Institut für deutsche Sprache, R 5, 6-13, D-68161 Mannheim, Germany.
Email:
brunner@ids-mannheim.de

1 Introduction

Speech, thought, and writing representation (ST&WR) is central to narrative theory, as it is important for constructing a fictional character and sheds light on the narrator–character relationship and the narrator's stance. The favored techniques not only vary between authors and genres, but have changed and developed over the course of literary history. An automated annotation of ST&WR would be valuable, as it could quickly deal with a large number of texts and allow a narratologist to study regularities and differences between different periods, genres, or authors.

This article presents the main results of a project that explored ways to recognize and classify this narrative device automatically, using surface information and methods of computational linguistics. The focus is not on building a specific application, but on discovering the possibilities and limitations of ultimately surface-based automatic annotation and its relationship to a manual annotation, which was generated with narratological concepts in mind.

In narratological theory, different systems for describing ST&WR have been developed. The categories used in the project are similar to those defined by Genette (1980) or Leech and Short (2007). Four types are considered: direct representation ('He thought: "I am hungry."'), free indirect representation, which takes characteristics of the character's voice as well as the narrator's ('Well, where would he get something to eat now?'),

indirect representation ('He said that he was hungry.') and reported representation, which can be a mere mentioning of a speech, thought, or writing act ('They talked about lunch.'). The goal of the project was to recognize and distinguish these types in German narrative texts.

Recognizing ST&WR implies an attempt to parametrize a narrative device to handle it with quantitative methods. This sets this project apart from other digital approaches to literary texts that often operate purely on a vocabulary level and are more focussed on thematic issues or on author or group-specific stylistics.

Other recent approaches to the automatic identification of ST&WR are usually not interested in the narrative techniques themselves. The annotation is just a step for some other goal, like identifying second-hand information (cp. Krestel *et al.*, 2008) or extracting a network of interrelations of fictional characters (cp. Elson and McKeown, 2010). Therefore, they tend to focus on the most frequent and unambiguous ST&WR techniques. This is usually direct ST&WR (e.g. Mamede and Chaleira, 2004; Elson and McKeown, 2010), and sometimes indirect ST&WR as well (e.g. Krestel *et al.*, 2008; Sarmento and Nunes, 2009). None of the cited approaches are for the German language, and only the first two deal with fiction, while the others are concerned with newspaper texts. Also, only rule-based methods are used to detect ST&WR, while in the project presented in this article, machine learning was tested as well.

This article first gives a general overview of the project, including a description of the corpus used as the basis for the research and an outline of the two approaches to detect the narratological feature, rule-based strategies, and machine learning. It then provides a detailed evaluation of the results for the four types of ST&WR and concludes with some additional remarks regarding those results.

2 The corpus

Basis for the research is a corpus containing thirteen short German narratives written between 1787 and 1913 (~57,000 tokens). The corpus has been manually annotated by the author with a set of twelve ST&WR categories adapted from narratological theory: direct, free indirect, indirect, and reported representation of speech, thought, or writing. Recognizing and classifying ST&WR can be a hard task even for a human annotator, especially when thought representation is concerned. To account for that, borderline cases were marked and categorized when doing the manual annotation. Such cases include, for example, non-factual ST&WR (e.g. 'He won't say that he is hungry.') or instances where there is no prototypical speech, thought, or writing act involved. For example, the sentence 'He was called Bill' uses a speech verb, but is pragmatically a paraphrase of 'His name was Bill', so it is classified as a borderline case of reported speech representation. The phrase 'She knew where he was' can be considered a borderline case of indirect thought representation, as 'to know' refers to a state rather than a specific thought act.

The annotation of the corpus is comparable with the annotation project conducted by Semino and Short for a corpus of English literary, autobiographical and newspaper texts (Semino and Short, 2004), but is something that has never been done for German literary texts before. The manual annotation not only gives empirical insight into the surface structures and the complex nature of ST&WR, but also serves as training material for machine-learning approaches and, most importantly, as reference for evaluation of the automatic recognizers. For the results presented in this article, only the distinction of the four ST&WR techniques direct, free indirect, direct, and reported was considered. Also, a broad definition of ST&WR was used, which included all borderline cases.

The corpus was compiled from public domain texts in a basic TEI format, which were downloaded from the site of the project TextGrid (http://www.textgrid.de/Digitale-Bibliothek, cp. Neuroth *et al.*, 2011). The intention was to give a good sample of older German narratives, and as you can see in Table 1 the corpus is diverse: it contains works of female as well as male authors, written in first as well as third person perspective, and the texts span a period of >100 years. They also have different narrative styles: Some contain a lot of dialogue (e.g. 'May', 'Janitschek'), while others are mostly

Table 1 Texts of the corpus

Year	Author	Title	Perspective	Gender	Tokens
1787	Musäus, J. K. A.	Die Entführung	3rd	m	5,222
1788	Bürger, G. A.	Münchhausen (1st Chapter)	1st	m	1,660
1802	Bernhardi, S.	Belinde	3rd	f	4,696
1805	Günderrode, K.	Geschichte eines Braminen	1st (3rd)	f	4,393
1807	Kleist, H. v.	Das Erdbeben in Chili	3rd	m	6,577
1812	Tieck, L.	Der blonde Eckbert	3rd (1st)	m	7,593
1825	Hauff, W.	Die Geschichte von Kalif Storch	3rd	m	4,741
1849	Hebbel, F.	Die Kuh	3rd	m	2,081
1878	May, K.	Die verwünschte Ziege	3rd	m	5,831
1889	Schnitzler, A.	Mein Freund Ypsilon	1st	m	4,976
1902	Janitschek, M.	Darüber kommt kein Weib hinweg	3rd	f	1,754
1913	Heym, G.	Der Irre	3rd	m	5,653
1913	Kafka, F.	Der Jäger Gracchus	3rd	m	2,045

descriptive (e.g. 'Musäus', 'Günderrode') or focus on representing the consciousness of the protagonist ('Heym'). Although most of the texts are modernized to some degree, there is also variety in spelling and punctuation. Because of this, the corpus is a fairly difficult input for an automatic recognizer, and there is a chance that the results would improve if more homogenous data were used.

Fig. 1 shows how frequent the different types of ST&WR are in the corpus according to the manual annotation. Each sentence containing an instance of ST&WR was counted for the respective type. If a sentence contained several types of ST&WR, which is often the case for indirect and reported ST&WR, it was counted multiple times. As a reference, the number of sentences that do not contain any type of ST&WR is also given. You can see that direct ST&WR is by far the most frequent type. Indirect and reported ST&WR are less frequent but still present in every text. Free indirect ST&WR is not well represented in the corpus owing to the fact that it is a technique favored by more modern works. It is present in only five of the thirteen texts, and most instances are concentrated in a single text.

3 Strategies for recognizing ST&WR

3.1 Tools and preprocessing

The rule-based components of the automatic recognizer are modular and realized as working prototypes in the General Architecture for Text Engineering (GATE) framework (http://gate.ac.uk). Machine learning was performed in R (http://www.rproject.org), using the randomForest package (cp. Liaw and Wiener, 2002). Preprocessing of the corpus included tokenization and sentence splitting, performed by tools that are part of the information extraction system ANNIE distributed with GATE (cp. Cunningham *et al.*, 2002). Morphological tagging was done with the German versions of the TreeTagger (Schmid, 1995) and the RFTagger (Schmid and Laws, 2008). The results of the RFTagger were only used when detailed morphological information was needed, e.g. for recognizing subjunctive mood.

3.2 Rule-based approach

For the rule-based approach, simple and robust methods were favored, which do not require advanced syntactic or semantic preprocessing, automatic reasoning, or complex knowledge bases. The modules make use of conventions like punctuation, and lexical and structural characteristics for different types of ST&WR. A central feature is a list of words that signal ST&WR, e.g. 'sagen (to say)' or 'flüstern (to whisper)'. As there are different indicators for each of the four different types, each has to be handled separately in the rule-based approach.

For direct representation, typographical features like quotation marks are central. However, though quotation marks can be good indicators, experiences with real texts showed that they are not as

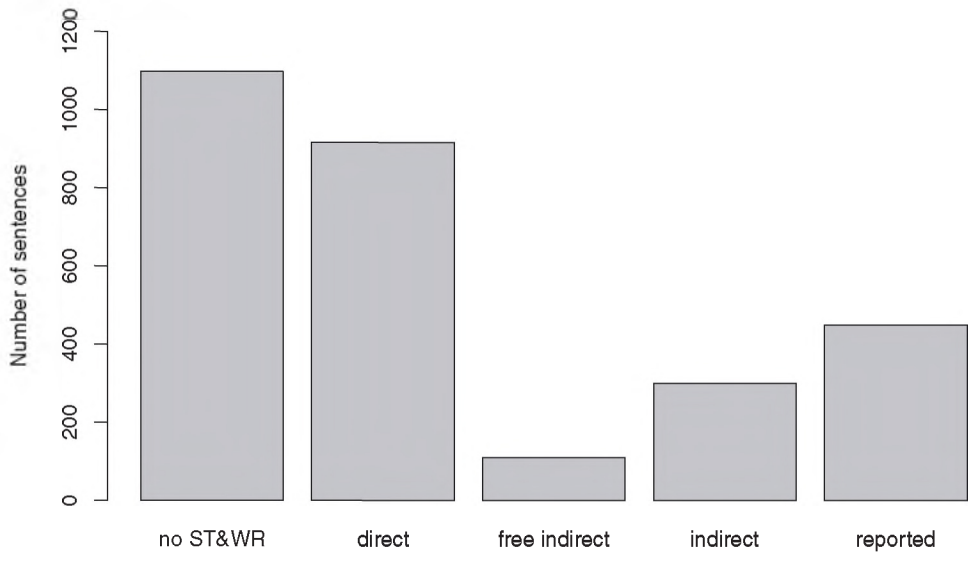

Fig. 1 Frequency of the different types of ST&WR in the corpus.

reliable as might be expected: quotation marks may have other functions (e.g. indicating irony), there are different conventions of usage, and they are certainly not obligatory for direct ST&WR. In the corpus, one text did not use quotation marks at all, while others skipped them for nested ST&WR or thought representation. It is not trivial to anticipate and fix such instances, as they depend on the preferences of the individual author or publisher, and especially for older non-standardized texts, punctuation can be idiosyncratic. Also, a lot of typeset and encoding errors are related to quotation marks. This particular problem did not arise in the corpus, but is something to bear in mind when using quotation marks as indicators. For these reasons, the rule-based recognition is supplemented by the detection of framing phrases, e.g. 'er sagte: (he said:)'. This ensures that the recognizer can find likely instances of direct representation even when no quotation marks are present in a text.

Indirect representation is the type of ST&WR that is most strongly defined by its structure. It is well described in linguistic research and has a distinct form: framing clause plus dependent clause. Patterns of surfaces and morphological categories are used to match the most typical realizations (e.g. 'Er sagte, dass er hungrig sei. [He said that he was hungry.]': signal word–followed by comma–followed by a specific conjunction–followed by a free match, concluded by a verb). Rules for three types of dependent clauses have been implemented: subordinate clause with a conjunction ('Sie fragt, ob es regnet. [She asks whether is rains.]'), infinitive clause ('Sie befahlen uns zu kommen. [They ordered us to come.]'), and subordinate clause in subjunctive mode but without conjunction ('Er sagte, er sei müde. [He said he was tired.]').

Reported representation is detected on a lexical basis and by making use of the results of the other recognizers. First, all possible indicators are annotated with the list of signal words. Then, all hits that are part of indirect ST&WR or of framing phrases of direct ST&WR are eliminated. The remaining signal words are assumed to indicate reported ST&WR.

Free indirect ST&WR is the hardest type to identify with a rule-based approach, as it does not have indicators, which are at the same time frequent and easy to detect on the text surface. For this type of ST&WR, sentences are assigned positive and negative points depending on whether they have characteristics that make free indirect ST&WR likely or unlikely. If a threshold is passed, the sentence is categorized as free indirect ST&WR.

4 Machine-learning approach

The machine-learning approach uses the Random Forest learning algorithm (Breiman, 2001) trained on the manually annotated corpus. Random Forest was chosen because it is considered one of the most accurate machine-learning algorithms to date (cf. Caruana *et al.*, 2008) and, as an ensemble learner, deals well with small training samples (cf. Polikar, 2006).

Eighty features were used. Some of those encoded general characteristics like sentence length or the percentage of different morphological categories (based on the annotation provided by the TreeTagger). Others were chosen based on assumptions about the characteristics of ST&WR types, e.g. attributes that encode the presence of verbs from the list of signal words or attributes encoding the use of first and second person pronouns, which may indicate direct ST&WR.

For each ST&WR type, a separate model is trained on positive examples (sentences containing the ST&WR type) and negative examples (sentences that do not contain the ST&WR type in question, but may well contain instances of other types). As Fig. 1 shows, positive examples are significantly less frequent than negative examples for all types except direct ST&WR. Therefore, resampling was used in the training phase: the positive examples were duplicated until their number matched that of the negative examples, thereby negating the bias of the learning algorithm toward favoring the dominant class.[1]

Instances used for machine learning were usually sentences. This means that the algorithm attempts to learn which sentences contain a certain type of ST&WR, but not what the exact boundaries are. However, instances of indirect and reported ST&WR are often shorter than a sentence, so many characteristics of the whole sentence might be irrelevant for recognizing them. To find out whether a tighter focus can improve the results, machine learning for these two types was also performed on shorter units, called 'sections'. These sections are created by splitting sentences after certain punctuation marks (comma, colon, semicolon) and before the word 'und [and]', but only if the resulting chunk still contains a verb. This usually ensures that coordinated nominal phrases are not split. The result is chunks that roughly correspond to grammatical clauses.

As the manual annotation of the corpus was needed as training material, the results presented below were gained by performing a tenfold stratified cross validation on the whole corpus: The corpus was divided into ten stratified parts. Nine parts were recombined and resampled to train a Random Forest model, which was then used to classify the instances of the tenth part. This process was repeated with a different part set aside as test data each time until all instances of the corpus were classified.

In the following sections, a detailed evaluation will be given for the four types of ST&WR. In each case, rule-based recognition as well as recognition based on machine learning was performed. It turned out that the two recognizers often annotated different instances, so it was tested whether combining the results leads to improvements. In one case, only instances found by both methods were counted (intersection Rule/ML) and in the other case all instances were counted that were found by either of the methods (union Rule/ML). Machine learning was usually performed on sentences, but for indirect and reported ST&WR, results of the section-based machine learning are presented as well. However, for the combinations only the sentence-based machine-learning results were used.

5 Evaluation for recognizing direct ST&WR

Evaluation is an analytic task itself: results are not only dependent on the exact configuration of the recognizer modules, but there are also different aspects that can be measured.

The first aspect we look at is the accuracy of the results. 'Precision' shows which percentage of the automatically annotated instances is correct. 'Recall' is the percentage of manually annotated instances that were found by the automatic methods. An ideal recognizer would produce good results for both measures, but in reality there is usually a

trade-off between them. To get a measurement for the overall performance, the harmonic mean of precision and recall is used, called 'F1 score'. This measurement combines the two values equally. The results for precision, recall, and F1 score always lie between 0.0 and 1.0, 1.0 meaning perfect success. To estimate the overall success, these figures are calculated for the whole corpus, i.e. the maximum amount of data available.

However, not only the overall accuracy of the recognizer is important but also the reliability of the results. To get an estimate for this, the F1 score was also calculated for each of the thirteen texts individually. As each text is a complete work of fiction, the results can be used to get an idea how the recognizer performs on different works. The standard deviation of the results serves as an estimate of how stable the performance of the recognizer is.

Table 2 shows the results for direct ST&WR. All techniques achieve F1 scores of 0.84 and more for the whole corpus. When the results of rule-based and machine-learning approaches are combined, either precision or recall can be maximized: the intersection gives a precision of 0.97, the union gives a recall of 0.96. This is the best accuracy achieved for recognizing any type of ST&WR.

The mean of the F1 scores for the individual texts is the same for all methods, but the values for the standard deviation vary: the results of the machine-learning approach are more stable with a standard deviation of 0.188 compared with 0.222 for the rule-based approach. This is because the success of the rule-based method strongly depends on whether the direct ST&WR is marked consistently by quotation marks. If this is the case, rule-based methods can outperform machine learning in accuracy. If the quotation marks are missing, however, their results get much worse, though the recognition does not fail completely as framing phrases are used as indicators as well. The results of machine learning are much less dependent on quotation marks.

Another way of looking at the results of the automatic recognition is how well they serve to predict the percentage of ST&WR used in a text. This is a more lenient evaluation, as it does not take into account whether the correct instances were found. However, predicting the percentages is useful for studies that are concerned with how the relative frequency of ST&WR usage developed over time or in different genres.

Fig. 2 illustrates the relationship between the true and predicted percentages for direct ST&WR by plotting them as lines. The percentage according to the manual annotation is represented by the thick line. The lines produced by the different kinds of automatic recognizers should be as similar to it as possible.

Two different kinds of similarity have been considered in the evaluation. The first is the 'mean absolute error', which measures the accuracy of the estimate: for each text, the percentage of sentences containing ST&WR according to the manual annotation is subtracted from the percentage of sentences in which the automatic recognizer detected ST&WR. This difference is taken as an absolute value, so it is not important whether the automatic recognizer found more or less than the true percentage. The mean of these values is the mean absolute error. As with the F1 scores, the standard deviation can be calculated to estimate how stable the error rate is.

The second way of looking at the similarity of the percentages is by calculating 'Pearson's correlation coefficient' for the real percentage values and the ones produced by the automatic recognizer. This tells us how well the real percentages can be predicted using the percentages found by the automatic recognizers, the ideal value of the correlation being 1.0. Unlike the mean absolute error, correlation is not concerned with accuracy: a consistent error of 5% for each text would give a good correlation. If the error were only between 1 and 3% but changed from text to text, the correlation would be worse even though the mean absolute error would be lower.

The visualization in Fig. 2 and the values in Table 3 confirm that the rule-based recognizer gives good estimates for some texts, but has problems with other texts (especially 'Bernhardi', 'Günderrode', 'Kleist', 'Tieck') where it over- or underestimates the percentage of direct ST&WR considerably. These problems are related to quotation marks and result in a high mean absolute error

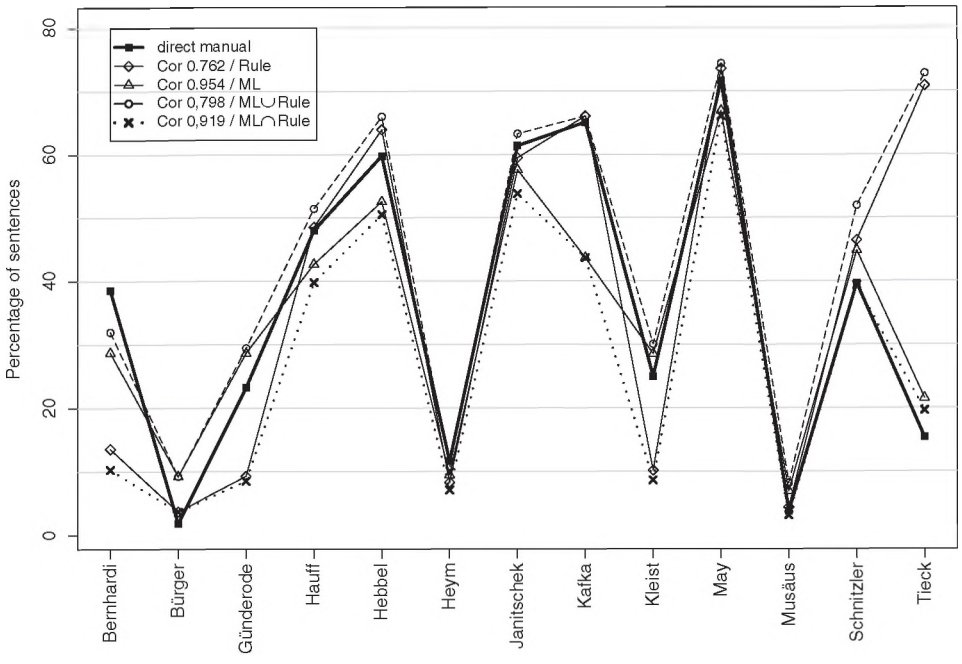

Fig. 2 Direct ST&WR: percentages.

Table 2 Direct ST&WR: measuring accuracy

Recognition method	Whole corpus			Individual texts	
	Precision	Recall	F1 score	Mean F1 score	Standard deviation
Rule-based	0.81	0.87	0.84	0.77	0.222
ML (sentence)	0.88	0.85	0.87	0.77	0.188
Intersection ML/Rule	0.97	0.77	0.86	0.77	0.191
Union ML/Rule	0.77	0.96	0.85	0.77	0.243

Table 3 Direct ST&WR: prediction of percentages

Recognition method	Mean absolute error	Standard deviation	Correlation
Rule-based	10.086	15.465	0.762
ML	6.553	4.888	0.954
Intersection ML/Rule	9.448	8.474	0.919
Union ML/Rule	8.876	14.874	0.798

of 10.086 with a standard deviation of 15.465 and a low correlation of 0.762. The machine-learning approach proves more accurate and more stable and also gives a much better correlation—for this corpus it is the best choice for predicting the percentages of direct ST&WR.

The experiments show that when recognizing direct ST&WR it is important to consider the status of quotation marks. If direct ST&WR is marked consistently, there is a good chance that the rule-based methods give accurate results. If this can not be ensured, machine learning is most likely the safer alternative.

6 Evaluation for free indirect ST&WR

As mentioned above, free-indirect ST&WR has no indicators that are at the same time frequent and easy to identify, so it is hard to capture with rules alone. The rule-based approach strongly relies on heuristics as well.

The figures in Table 4 confirm that for such an elusive technique, machine-learning methods are clearly superior, as they can pick up on structural indicators, which are not obvious for humans. The F1 score of 4.0 achieved by the

Table 4 Free indirect ST&WR: accuracy

Recognition method	Whole corpus		
	Precision	Recall	F1 score
Rule-based	0.24	0.44	0.31
ML (sentence)	0.63	0.29	0.40
Intersection ML/Rule	0.68	0.19	0.30
Union ML/Rule	0.26	0.54	0.35

Table 5 Free indirect ST&WR: prediction of percentages

Recognition method	Mean absolute error	Standard deviation	Correlation
Rule-based	5.062	2.748	0.768
ML (sentence)	1.681	3.598	0.973
Intersection ML/Rule	2.107	4.763	0.849
Union ML/Rule	4.685	3.108	0.935

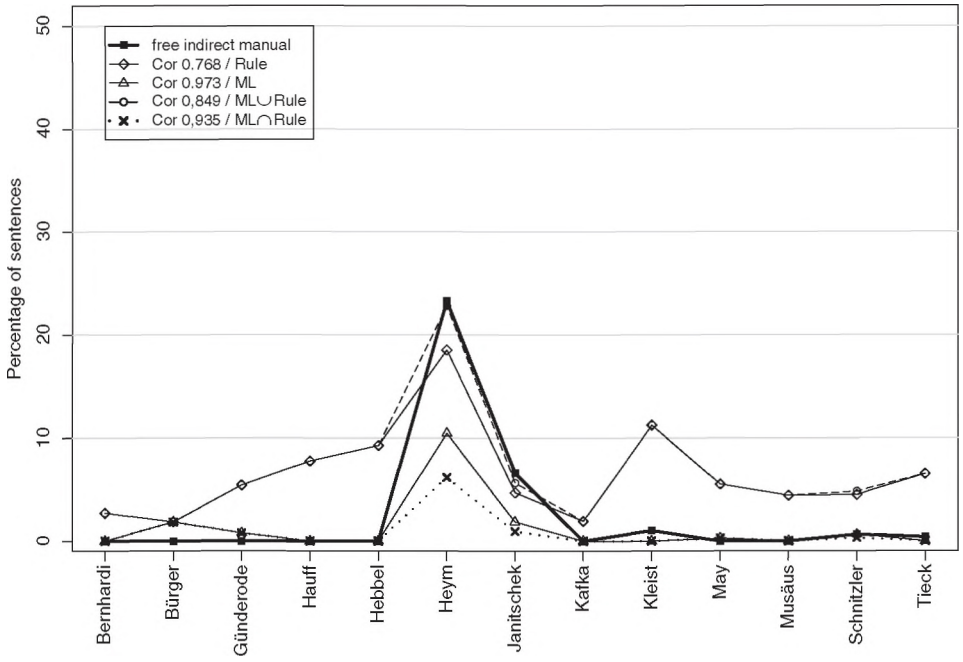

Fig. 3 Free indirect ST&WR: percentages.

machine-learning approach is much lower than the scores for direct ST&WR, but considerably better than the F1 score achieved by rule-based recognition. The precision of 0.63 is good, but the recall is low. This is a general problem related to the resampling method used in the machine-learning process: if the training data are imbalanced—as is the case for free indirect ST&WR—the instances of the less-frequent class need to be duplicated many times and the resulting model tends to strongly favor precision over recall when classifying new instances.

For this ST&WR technique, the mean of the F1 scores for the individual texts and the standard deviation were not calculated. This is because eight of the thirteen texts do not contain instances of free indirect ST&WR at all. The F1 scores of such texts can only be 1.0, if the automatic recognizer does not detect an instance, or 0.0, if it does detect any number of instances. These values are too rough and extreme and make the mean and standard deviation meaningless.

Fig. 3 gives the percentages of free indirect ST&WR in the individual texts and shows the uneven distribution where nearly all instances are concentrated in the text 'Heym'. The machine-learning method captures this distribution rather well, while the rule-based recognizer finds a similar

Table 6 Indirect ST&WR: accuracy

Recognition method	Whole corpus			Individual texts	
	Precision	Recall	F1 score	Mean F1 score	Standard deviation
Rule-based	0.81	0.62	0.71	0.68	0.116
ML (sentence)	0.62	0.47	0.53	0.46	0.223
ML (section)	0.38	0.70	0.50	0.47	0.115
Intersection ML/Rule	0.85	0.38	0.53	0.46	0.228
Union ML/Rule	0.66	0.71	0.68	0.64	0.151

Table 7 Indirect ST&WR: prediction of percentages

Recognition method	Mean absolute error	Standard deviation	Correlation
Rule-based	3.850	3.800	0.848
ML sentence	4.000	3.318	0.865
ML section	9.982	3.323	0.943
Intersection ML/Rule	7.537	4.667	0.832
Union ML/Rule	2.653	2.089	0.880

percentage of free indirect ST&WR in several texts beside 'Heym'.

Table 5 shows that consequently, machine learning gives a much lower mean absolute error and a much better correlation than the rule-based method.

Free indirect ST&WR is the most infrequent technique in the corpus (see Fig. 1), so the results are least reliable. As so many instances are from the same text, there is also a danger that the machine-learning approach picked up characteristics of the author style instead of the ST&WR type. Still, the results indicate that detecting free indirect ST&WR with machine learning is a strategy worth pursuing in further research.

7 Evaluation for indirect ST&WR

Recognizing indirect ST&WR achieves the second best results overall. Table 6 shows that rule-based methods give the best accuracy: an F1 score of 0.71. The precision of 0.81 is only exceeded by the results for direct ST&WR. The standard deviation of the F1 scores for individual texts is 0.116 and shows that the rule-based method also produces fairly stable results. The machine-learning approach—both sentence-based and section-based—only gives F1 scores of 0.50–0.53, and sentence-based learning has a standard deviation of 0.223, about twice as high as the rule-based approach. It is not surprising that indirect ST&WR favors the rule-based approach, as it has a distinct prototypical structure that can be captured fairly well by pattern rules.

When looking at the percentages in Fig. 3, you can see that the lines for rule-based recognition and the sentence-based machine-learning approach are similar. The recognizers have similar problems as well, especially underestimating the percentage of indirect representation in the texts 'Bürger' and 'Günderrode'. However, this does not mean that the two methods always predict the same instances. This is apparent because the lines for the combinations—intersection and union—are distinct from each other.

Table 7 shows that the mean absolute error and its standard deviation are also similar for the rule-based recognition and the sentence-based machine-learning approach. Just considering those measurements, union ML/Rule seems the best choice with a mean absolute error of 2.653 and a standard deviation of 2.089.

However, when looking at the correlation, section-based machine learning gives an interesting result: this method has a high mean absolute error of nearly 10, but a much better correlation than any other method: 0.943. Even though it strongly overestimates the percentage of indirect ST&WR, it does so consistently for each text. So the results can actually be more useful for predicting how the true percentage changed between texts than the other, more exact recognizers.

This demonstrates that the best strategy for recognizing ST&WR may also depend on what is prioritized in the evaluation. For indirect representation, rule-based methods are clearly superior when it comes to accuracy, but machine learning based on sections, even though it gives inaccurate results, seems most suited for comparing percentages.

8 Evaluation for reported ST&WR

The rule-based and the machine-learning approach are more or less tied when it comes to recognizing reported ST&WR. Table 8 shows that the best F1

Table 8 Reported ST&WR: accuracy

Recognition method	Whole corpus			Individual texts	
	Precision	Recall	F1 score	Mean F1 score	Standard deviation
Rule-based	0.51	0.64	0.57	0.57	0.111
ML (sentence)	0.56	0.45	0.50	0.43	0.210
ML (section)	0.49	0.54	0.52	0.48	0.098
Intersection ML/Rule	0.63	0.37	0.47	0.40	0.182
Union ML/Rule	0.49	0.71	0.58	0.56	0.137

Table 9 Reported ST&WR: Prediction of percentages

Recognition method	Mean absolute error	Standard deviation	Correlation
Rule-based	5.290	4.100	0.849
ML sentence	3.657	2.653	0.954
ML section	3.368	2.252	0.942
Intersection ML/Rule	7.514	3.782	0.938
Union ML/Rule	8.386	4.609	0.897

score, 0.58, can be achieved by the union of both approaches. However, rule-based methods alone are almost as successful and have a somewhat lower standard deviation.

When looking at the prediction of percentages, however (Fig. 5 and Table 9), machine-learning methods are preferable: they give the lowest mean absolute error, lowest standard deviation, and the best correlation. In this case, sentence-based and section-based machine learning produce similar results, and the section-based machine learning does not heavily overestimate the percentages like it does for indirect ST&WR.

So you see a similar trend as for indirect ST&WR: rule-based methods are better for accuracy, but machine learning is better for predicting percentages. However, the trend is less clear, as the overall results are more similar. It should also be noted that the rule-based recognition of reported ST&WR relies on the results of the recognition of direct and indirect ST&WR, which are used to eliminate indicators. The machine-learning recognizer stands alone. It is possible that its results could be improved if it made use of the results of other recognizers as well.

9 Conclusion

The results of the project show that automatic recognition of ST&WR is not easy, but not a hopeless task either. The best results could be achieved for direct ST&WR, followed by indirect and then reported ST&WR. Free indirect ST&WR proved to be most difficult, but even for this type there was some success.

Rule-based strategies are most successful for ST&WR types with clear patterns and conventions, like indirect ST&WR and consistently marked direct ST&WR. They also tend to give the most accurate results (except for free indirect ST&WR). Recognizers based on machine learning are especially useful for elusive types like free indirect ST&WR oder unmarked direct ST&WR. They tend to be somewhat less accurate, but more stable, and are especially good when it comes to correlation, where they outperformed rule-based methods every time. So both the rule-based and the machine-learning approach have their merits and there is no clear favorite. Combining the results of the two approaches can be used to maximize either precision (intersection) or recall (union). However, the combinations did not lead to striking improvements, so perhaps a more sophisticated way of merging the rule-based and the machine-learning approach is called for to get the best of both worlds. The choice between the different methods and their configuration also strongly depends on what kinds of results are needed in the individual research project that uses a ST&WR recognizer.

There are two things to keep in mind when comparing the two approaches based on the results of this project. First, as the results presented here are based on stratified cross validation, the machine-learning methods were tested on data that were similar to their training data. Although the texts of the corpus themselves are diverse, it is not clear how well the machine-learning models trained on these data perform for different input. When using the machine-learning approach, it would be best to train the learner on a sample of the corpus that will be used in the actual application. Second, the

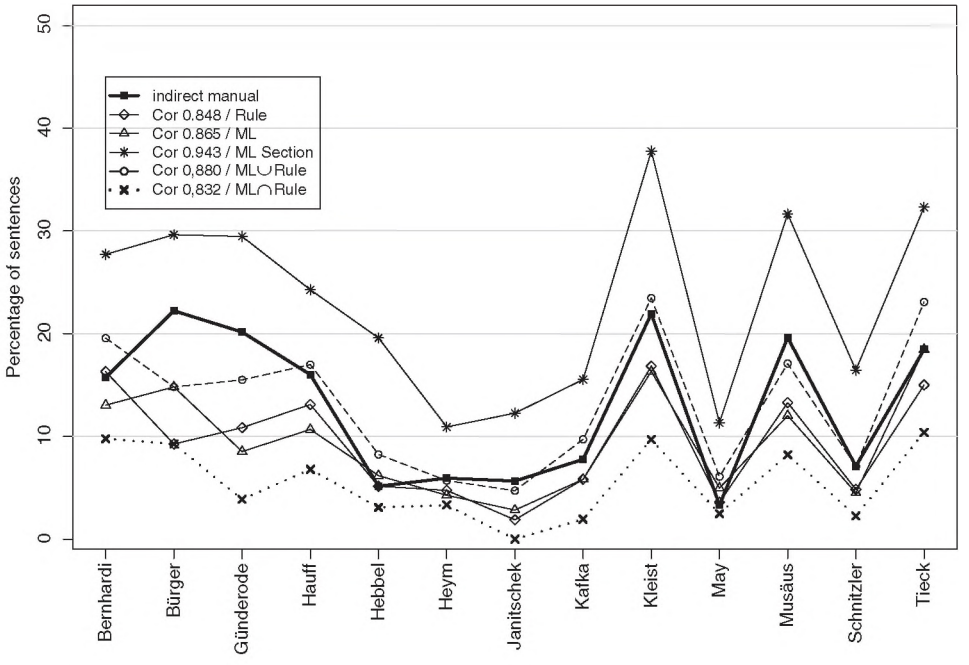

Fig. 4 Indirect ST&WR: percentages.

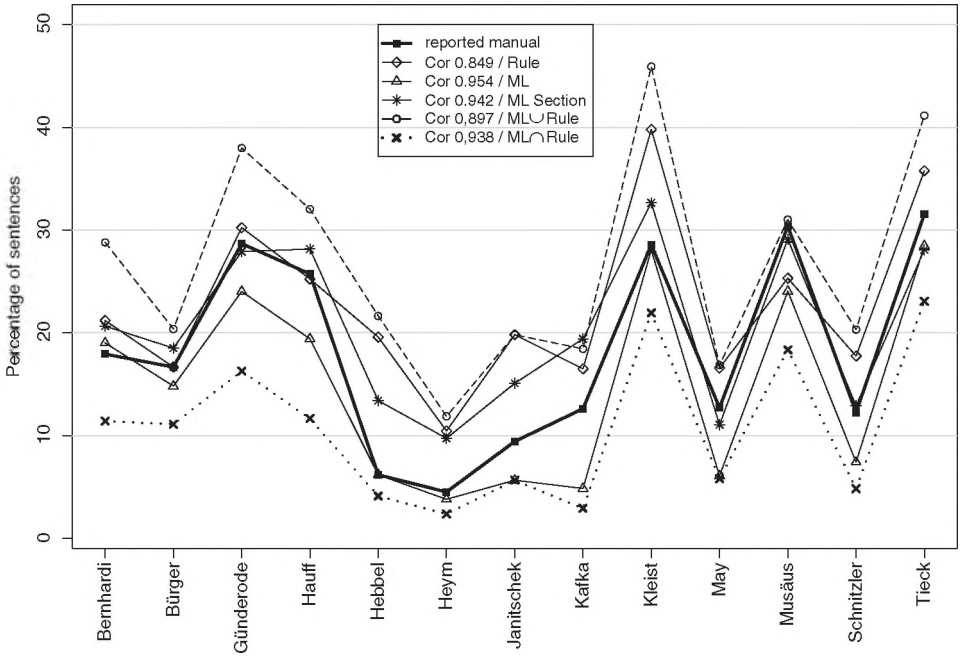

Fig. 5 Reported ST&WR: percentages.

rule-based methods were basic, using mostly lexical indicators and surface-based pattern rules.[2] There is definitely room for refinement, especially for indirect and reported representation, but it would be a time-consuming task most likely involving advanced preprocessing and it is unclear whether the improvement in performance would justify this investment.

As mentioned above, the manual corpus annotations also marked borderline cases of ST&WR. The results presented above are based on a rather broad definition of ST&WR, which included all of these cases as positive examples for ST&WR. Tests were done to check how the performance of the automatic recognition is influenced by a stricter definition of ST&WR. It turned out that for both the rule-based and the machine-learning approach the results get worse the more restricted the definition becomes. For the rule-based methods, this is in part due to the fact that they were developed with the broad definition in mind and would need specific rules for additional restrictions. The machine learning suffers greatly from the smaller number of positive training examples. On the whole, it seems an additional difficulty to distinguish between what is considered prototypical ST&WR and what is not.

The results of this project give hope that it is not impossible to tackle the task of automatically detecting a narrative feature like ST&WR and hopefully represent some steps on the way. Although it was not part of the project to develop a stable and easy-to-use recognizer for ST&WR, the GATE modules developed for rule-based recognition and the R scripts used for machine learning are working prototypes and will be made available under a creative commons license. The same is true for the manually annotated corpus. If you are interested in using any of those components for your own research, please contact the author.

References

Breiman, L. (2001). Random forests. *Machine Learning*, **45**(1): 5–32.

Brunner, A. (2012). Automatic recognition of speech, thought and writing representation in German narrative texts. *Digital Humanities 2012: Conference Abstracts*. Hamburg: Hamburg University Press, pp. 135–6.

Caruana, R., Karampatziakis, N., and Yessenalina, A. (2008). *An empirical evaluation of supervised learning in high dimensions, Proceedings of the 25th Annual International Conference on Machine Learning (ICML 2008)*. Helsinki.

Cunningham, H., Maynard, D., Bontcheva, K., and Tablan, V. (2002). *GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications, Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*. Philadelphia, PA.

Domingos, P. (1999). *MetaCost: A General Method for Making Classifiers Cost-Sensitive, Proceedings of the Fifth International Conference on Knowledge Discovery and Data Mining*. San Diego, CA, USA.

Elson, D. K. and McKeown, K. R. (2010). *Automatic Attribution of Quoted Speech in Literary Narrative, Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence (AAAI-10)*. Atlanta, GA.

Genette, G. (1980). *Narrative Discourse. An Essay in Method*. Oxford: Blackwell.

Krestel, R., Bergler, S., and Witte, R. (2008). *Minding the Source: Automatic Tagging of Reported Speech in Newspaper Articles. In E. L. R. A. (ELRA), Proceedings of the Sixth International Language Resources and Evaluation Conference (LREC 2008)*. Marrakech.

Leech, G. and Short, M. (2007). *Style in Fiction. A Linguistic Introduction to English Fictional Prose 2*. London: Pearson Education Limited.

Liaw, A. and Wiener, M. (2002). Classification and Regression by randomForest. *R News*, **2**(3): 18–22.

Mamede, N. and Chaleira, P. (2004). Character Identification in Children Stories. *EsTAL 2004 – Advances in Natural Language Processing, LNCS*. Berlin/Heidelberg: Springer, pp. 82–90.

Neuroth, H., Lohmeier, F., and Smith, K. M. (2011). TextGrid—virtual research environment for the humanities. *The International Journal of Digital Curation*, **6**(2): 222–31.

Polikar, R. (2006). Ensemble based systems in decision making. *IEEE Circuits and Systems Magazine*, **6**(3): 21–45.

Sarmento, L. and Nunes, S. (2009). *Automatic Extraction of Quotes and Topics from News Feeds, Proceedings of DSIE'09 - 4th Doctoral Symposium of Informatics Engineering*. Porto.

Schmid, H. (1995). *Improvements on Part-of-Speech Tagging with an Application to German, Proceedings of the ACL SIGDAT-Workshop*. Dublin.

Schmid, H. and Laws, F. (2008). *Estimation of Conditional Probabilities With Decision Trees and an Application to Fine-Grained POS Tagging*. Manchester, UK: Coling.

Semino, E. and Short, M. (2004). *Corpus Stylistics. Speech, Writing and Thought Presentation in a Corpus of English Writing*. London/New York: Routledge.

Notes

1 In the outline of this project presented in the abstract for Digital Humanities 2012 (Brunner, 2012), the problem of the unbalanced data set was solved in a different way: combining Random Forest with the meta classifier MetaCost (Domingos, 1999) made the learning cost sensitive. A misclassification of a positive instance was penalized by assigning higher costs. The general trend of the results is similar, except that resampling favors precision over recall, while cost-sensitive learning is more balanced in this respect. Which behavior is preferable depends on the intended application of the recognizer. For the final evaluation presented in this article, resampling was chosen mainly for pragmatic reasons, as no stable implementation of MetaCost was available for the R environment at the time the project was done.

2 The level of complexity of the rule-based approaches to recognizing direct and indirect ST&WR that were cited in the introduction is equal or somewhat higher. Krestel *et al.* for example include more sophisticated argument structures for speech verbs (cf. Krestel *et al.*, 2008), but in general similar indicators are used, especially quotation marks and signal words.